Combating false information in Emergencies: Leveraging LLMs for Targeted False Information and Clarification Generation

Jiahong Lin *

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: kksmi18@163.com

Abstract. The rapid dissemination of false information during emergencies has become a significant challenge in managing online public opinion. While large language models (LLMs) have enhanced the speed and efficiency of information generation, they have also exacerbated the complexity of public sentiment by facilitating the spread of both false and clarifying information. This paper addresses the critical need for targeted clarification information to counteract false narratives during emergencies. We propose a novel approach by fine-tuning open-source LLMs to generate both false information and corresponding clarification texts, tailored to specific emergency scenarios and public opinion dynamics. By constructing a high-quality dataset of 1,715 paired false and clarification information samples from authoritative platforms, we employ a task-separated fine-tuning strategy using LoRA (Low-Rank Adaptation) to optimize model performance. Our evaluation metrics, including text fluency (BLEU), novelty (NOV), and diversity (DIV), demonstrate that fine-tuned models, particularly LLaMA3.1, excel in generating coherent and relevant texts. The results highlight the potential of LLMs in both generating and debunking false information, offering a robust framework for improving public opinion management during emergencies. This research contributes to the growing body of work on false information mitigation and provides practical insights for leveraging LLMs in crisis communication.

Keywords: Emergency Events; Large Language Models; Information Generation; AI-generated content (AIGC).

1. Introduction

The spread of false information in the online public opinion during emergencies has been escalating. While the rise of large language models has facilitated the rapid generation and dissemination of information, it has also intensified the complexity of public opinion. The competitive dissemination of false and clarifying information has become increasingly intense, presenting numerous challenges to society. In the field of emergency event-related information generation research, most studies focus solely on the credibility of generated content, without delving into the mechanisms behind the generation process and the corresponding clarification information. This leaves them ill-equipped to address the growing complexity of false information generation under large language models.

From the perspective of information generation, much of the false information during emergencies originates from malicious fabrication by certain self-media outlets seeking to gain traffic. With the rapid development of AI-generated content (AIGC) technology, more and more rumor-mongers are exploiting the powerful capabilities of generative models to quickly create large amounts of false information. This severely disrupts the stability that should be maintained during emergencies. Therefore, it is crucial to study the intrinsic mechanisms behind the generation of false information by generative models. By analyzing this mechanism, we can explore how to use relevant technological means to generate targeted clarification information, which can be used to counter the spread of false information in real time and effectively curb the diffusion of its negative effects. This would help maintain the stability and harmony of the online public opinion environment in a timely manner.

By studying the generation of rumor and clarification information during emergency events, it will not only help relevant authorities understand potential rumors related to an event in advance and

prepare emergency management plans, but also enable the automated and efficient generation of highly correlated information to combat the spread of rumors. This would help prevent the situation from escalating beyond control.

2. Related works

Text generation, also known as natural language generation, is a type of text generation task and one of the most critical subfields in Natural Language Processing (NLP). Its goal is to produce credible and readable human language text from various forms of input data, including text, images, tables, and knowledge bases. With the advent of deep learning, neural network models have become the dominant approach for text generation, achieving significant success in generating natural language text. Deep neural generation models typically adopt a sequence-to-sequence framework based on an encoder-decoder architecture [1].

In recent years, pre-trained language models (PLMs) have seen remarkable advancements in the field of NLP[2]. The core idea is to first pre-train a model on large-scale unsupervised corpora and then fine-tune it on supervised downstream task datasets. With the emergence of Transformer[3] architectures and increased computational power, the design of PLMs has evolved from shallow to deeper architectures, exemplified by models like BERT [4] and OpenAI GPT [5]. Extensive studies have shown that PLMs can encode vast amounts of linguistic knowledge from pre-training corpora into their large-scale parameters and learn general and contextual representations of language through specially designed objectives, such as next-token prediction. Researchers have proposed leveraging PLMs for text generation tasks [6]. By pre-training on large-scale corpora, PLMs acquire the ability to accurately comprehend natural language and fluently express it in human language—both of which are essential capabilities for completing text generation tasks.

The task of generating information text for emergency events is a specialized text generation task tailored to specific scenarios and needs. Emergency event information encompasses descriptions of events, emergency response information, and online public opinion dynamics, including trends reflected through social media, news platforms, and public opinions. Insights into public perceptions, reactions, and anxieties about events are crucial for understanding societal emotional fluctuations. Among these, the spread of false information during emergencies has become a focal area of research, as such information tends to proliferate under conditions of information scarcity and heightened social emotions. This can lead to public panic, misguide emergency responses, and even jeopardize public safety. The superior text generation performance of large language models (LLMs)—a subset of PLMs characterized by extensive parameters and training on massive datasets—makes them a powerful tool for malicious actors aiming to flood online platforms and social media with content during emergencies to influence public perception.

False information spreads rapidly through social media, making online public opinion prone to spiraling out of control, while clarifying information often lags, struggling to mitigate its impact effectively. Detecting false information requires complex data analysis and technical support, further increasing the challenge of response. The threat of LLMs generating false information is among the most frequently cited risks as these models continue to advance [7,8]. Vykopal et al. [9] conducted a systematic study on the ability of large language models to generate false information during emergencies. They evaluated 10 mainstream LLMs across 20 different emergency topics for their capability to produce fake news articles aligned with harmful false narratives. Most LLMs were able to generate convincing news articles consistent with dangerous narratives, with differences among the models largely attributed to their parameter sizes. Larger models produced higher-quality text and exhibited a stronger tendency to agree with false narratives. Furthermore, providing additional context (e.g., event title and summary) increased the model's alignment with false narratives and enhanced the effectiveness of fabrication. However, current false information generated by these models can still be identified by mainstream rumor detection systems. Huang et al. [10] conducted an in-depth investigation into ChatGPT’s ability to generate false information, proposing four prompting

methods: altering the original meaning of the event text, fabricating stories, creating fictional content, and multi-step prompting. Their evaluation of generated sample quality revealed that ChatGPT's self-assessed accuracy was 72.5%, while human evaluation rated it at only 54.8%, indicating that the generated fake news is challenging for humans to detect. Sun et al. [11] proposed a novel fake news generation method called VLPrompt, which eliminates the need for manually collecting false information. This approach guides LLMs to generate false information using tailored prompts while maintaining contextual consistency and preserving details of the original text. They created a dataset, VLPAN, containing real news, manually written fake news, and LLM-generated fake news, for use in false information detection experiments. The results showed that VLPrompt significantly outperformed other methods, generating more realistic false information that was more effective at deceiving both humans and automated detectors.

These studies focus on leveraging the linguistic capabilities of large language models (LLMs) to explore the generation of false information, underscoring the significant potential of LLMs in generating false information during emergency events. However, these approaches primarily rely on prompt engineering to utilize general-purpose LLMs for generating prompts and have not specifically enhanced this capability by fine-tuning LLMs for false information generation.

In contrast, Lai et al. [12] developed a model named RumorLLM based on open-source LLMs, which fine-tunes the base models to generate rumor-like content. The RumorLLM method incorporates the distinctive writing styles and content features of rumors through fine-tuning. It utilizes techniques such as Porter stemming and TF-IDF to process text and extract key features, combining them with prompt engineering to guide the model in generating rumor-like text. This approach extracts keywords/stems from real rumors and reconstructs prompts using these features. The fine-tuned RumorLLM is then used to expand and generate new rumor-like texts, aiming to imitate the style and content of real rumors while introducing variations. This enables the generation of augmented fake news datasets to address class imbalance issues, helping balance datasets and improve the overall performance of detection models.

Although research on generating false information for emergency events is becoming increasingly mature, it remains in an exploratory phase. These studies primarily aim to leverage the superior generative performance of LLMs to optimize false information detection. However, they often overlook the importance of clarification and explanatory information in addressing false information during emergencies. The timely availability of clarification information can effectively combat the spread of false information, reduce the response time for public opinion emergency management, and mitigate the negative impact of false information dissemination.

Based on previous work on false information generation, this study fine-tunes open-source LLMs to generate false information texts that are better aligned with emergency scenarios, real-world situations, and public opinion dissemination needs. Additionally, it constructs a dataset of clarification and explanatory information tailored to specific rumors for fine-tuning, enabling the generation of targeted clarification and rumor-refuting texts.

3. Method

3.1. Model Selection

The public opinion information studied in this paper focuses on significant Chinese social network false information and corresponding true clarification information during emergency events. To better align with the characteristics of emergency event information and generate competitive true and false information effectively in Chinese, this study selects four prominent, high-performing, highly rated, and knowledge-rich open-source large language models (LLMs). These are:

- Qwen2-7B-Instruct, a 7-billion-parameter model from Alibaba [13],
- Baichuan2-7B-Chat, a 7-billion-parameter model from Baichuan Intelligence [14],
- GLM4-9B-Chat, a 9-billion-parameter model from Zhipu AI [15],
- LLaMA3.1-8B-Instruct, an 8-billion-parameter model from Meta [16].

These models excel in Chinese text generation and can precisely generate information tailored to specific requirements based on prompts. For brevity, they will be referred to as Qwen2, Baichuan2, GLM4, and LLaMA3.1 in subsequent sections.

Qwen2 is pre-trained on a 7-trillion-token multilingual dataset and utilizes innovative architectures like GQA and DCA, excelling in tasks involving code, mathematics, and multilingual capabilities.

GLM4 leverages a 10-trillion-token pre-training dataset with a focus on optimizing Chinese alignment and long-context processing. It approaches GPT-4 level performance and supports extensive tool usage.

Baichuan2 is pre-trained on a 2.6-trillion-token dataset, demonstrating outstanding performance in vertical domains such as healthcare and law, with significant improvements in mathematics and coding tasks.

LLaMA3.1 is pre-trained on a 15-trillion-token dataset, employing a simplified architecture and DPO training methodology, achieving near GPT-4 level performance in mathematics, reasoning, and code generation tasks.

According to OpenCompass evaluations, GLM4-9B-Chat ranks first in the Chinese domain, Qwen2 comes second due to its high-quality Chinese training dataset, LLaMA3.1, despite its extensive training, shows slightly weaker Chinese logical reasoning, and Baichuan2 excels in specific fields.

3.2. Dataset Construction

This study focuses on the text generation of false information and clarification information within public opinion during emergencies, with an emphasis on the correlation between false information and corresponding truthful information. Existing datasets, such as Weibo-16[17], WeChat[18], Weibo-20[19], Weibo-21[20], primarily focus on false information detection and lack paired relationships between false information and corresponding clarifications. Major events such as the COVID-19 pandemic and the Russia-Ukraine conflict highlight the importance of systematic rumor-refuting platforms.

In the Chinese internet context, platforms like Tencent Fact Checker (https://vp.fact.qq.com/home),The Paper Fact Check (https://www.thepapercheck.com/large), China Fact-Check (https://chinafactcheck.com/), and the China Internet Joint Clarification Center (https://www.piyao.org.cn/bq/index.htm) provide authoritative clarifications for false information in domains like healthcare, international events, and political-military topics. Using web scraping techniques, this study collected rumors and corresponding clarification articles from these platforms. After data cleaning and pairing processes, a high-quality dataset containing 1,715 pairs of false and clarification information was constructed.

To create fine-tuning instruction sets, Qwen2-40B-Instruct was selected due to its stability in Chinese generation tasks and relatively low resource requirements (approximately 44GB VRAM), considering GPU limitations. Prompt design guided the model to summarize and extract keywords from false information and clarification texts. API requests were employed to generate diverse results, and pattern matching was used to extract standardized JSON-formatted text, leading to an initial instruction dataset. Figure 1 shows the process of constructing a dataset for fine-tuning.

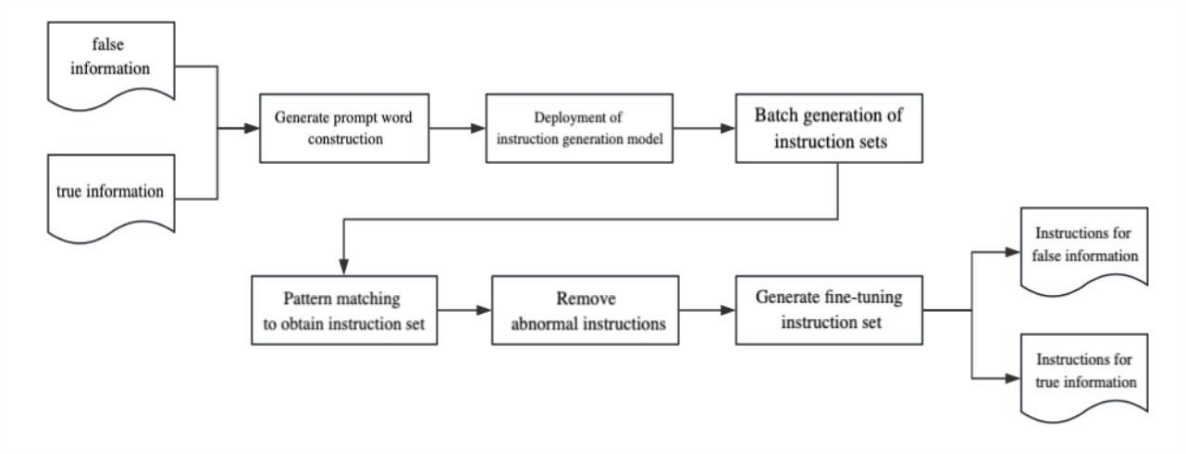


Fig 1. The process of constructing a dataset for fine-tuning.

The prompt design incorporated:

- Keywords relevant to emergency events,
- Information types (Politics, health care, social life, financial economy, emotional values, natural disasters),
- Text sentiment (positive, neutral, negative),
- Tone (Authoritative, inciting, intimidating, affinity, neutrality, inducement).

This ensured that the generated false information closely mimicked the characteristics of rumors in real-world social networks. After multiple independent inference runs and manual filtering to remove erroneous results, a final high-quality fine-tuning instruction set of 7,268 pairs (14,536 entries) was produced. This dataset will be used for subsequent model fine-tuning.

Table 1. Example of fine-tuned instruction set generation.

Item	Content
Prompt words	· Based on the following examples of rumor text and Clarification text, please generate 2 instruction data set entries for different tasks. Entries should relate directly to the given text content and propose relevant tasks. Rumor text: {text} Clarification text: {text2} Task 1: False information generation 1. Summarize the rumor text and Clarification text, and list 6 related keywords. 2. Based on these keywords, construct an instruction to generate false information. 3. The instruction should guide the generation of a piece of information with a specific type of false information (choose from politics, health care, social life, finance and economics, emotional values, natural disasters), specific emotional polarity (positive, neutral, negative) and a specific tone (authority, incitement, intimidation, affinity, neutrality, inducement) false information. 4. Use the following JSON format to generate entries: {{ "instruction": "Based on keywords [keyword 1, keyword 2, keyword 3, keyword 4, keyword 5, keyword 6], please generate an entry for [False information type] false information spread on social media uses [tone type] tone and reflects [emotional polarity] emotion. ", "input": "", "output": "The false information content is generated according to instruction requirements. " }} Task 2: Rumor-refuting information generation 1. Based on the Clarification text, construct an instruction to generate Clarification information. 2. Instructions should lead to a refutation or explanation of the false information generated in Task 1.

	3. Use the following JSON format to generate entries: {{ "instruction": "Please generate an authoritative refutation response to the following false information: [false information generated by task 1]", "input": "", "output": " Content of refutation information generated based on refutation text" }} Please ensure: 1. All generated content is directly related to the given text, and the output item is not empty. 2. The generated instruction string is a valid JSON format string that can be parsed. 3. The content is of high quality and accurate, with no redundant information except json format strings. .
Temperture	0.7
Number of Generations	5
Examples of false information	A forest fire broke out in Pengshui, Chongqing
Example of rumor refuting text	· On August 21, a netizen posted a video with the text "Peng Shui Baojia Sudden Forest Fire" in Chongqing, which attracted attention. After checking with the local government departments, there has been no such incident in Pengshui recently, and the information spread on the Internet is a rumor. At present, the rumor clues have been transferred to the relevant departments for further investigation and disposal of the rumor-mongers.
Example of False Information Generation Instruction	· { "instruction": "Based on the keywords [fire, pengshui, sudden, rumor-debunking information, rumors, cyberspace], please generate a [natural disaster] type of false information for social media communication, use a [intimidating] tone, and reflect [negative] emotions."} , "input": "", "output": "Urgent notice! The wildfire in Baojia Town, Pengshui County, broke out again, and the flames rose and spread to the town like a tornado, and the villagers fled in panic, and the scene was extremely tragic. All immediately away from the mountains! The fire is difficult to control, and the Internet is full of calls for help and crying, please be cautious and avoid dangerous areas! " }
Example of rumor refuting information generation	· { "instruction": "Please generate an authoritative response to the following false information: Urgent notice! The wildfire in Baojia Town, Pengshui County, broke out again, and the flames rose and spread to the town like a tornado, and the villagers fled in panic, and the scene was extremely tragic. All immediately away from the mountains! The fire is difficult to control, and the Internet is full of calls for help and crying, please be cautious and avoid dangerous areas! ", "input": "", "output": "Statement to refute rumors: Recently, false information has been spread through social media that the mountain fire in Baojia Town, Pengshui is critical. After strict verification, there have been no fires or similar dangers in our county; At present, everything is operating as usual, please do not panic or spread false information to mislead others, and jointly maintain the authenticity and tranquility of the online environment. Any incident or news should be subject to the public announcement of the government and professional institutions! " },

3.3. The process of fine-tuning

This study employs LoRA (Low-Rank Adaptation) for efficient instruction fine-tuning of open-source large language models. Experimental results revealed that mixing instruction sets for false information and clarification information during fine-tuning caused unstable model responses and logical inconsistencies. Therefore, a task-separated fine-tuning strategy was adopted. This involved fine-tuning the model separately for false information and clarification tasks, followed by merging the best checkpoint weights to ensure stable responses across tasks while avoiding quantization to minimize precision loss. The specific process of fine-tuning is shown in Fig. 2.

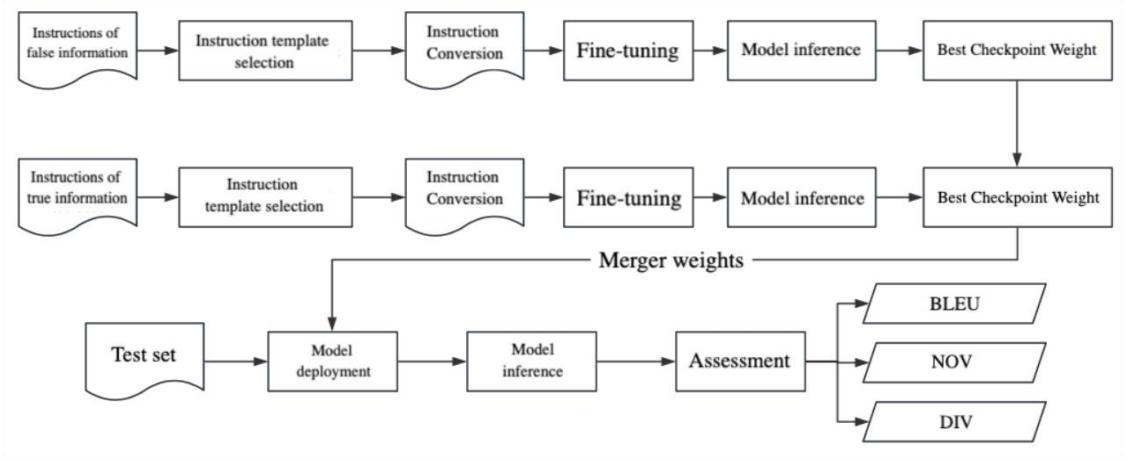


Fig 2. The specific process of fine-tuning.

For evaluation, 1,000 instructions were randomly sampled from each instruction set as test datasets. LoRA fine-tuning was conducted using half-precision floating point (FP16) to reduce memory consumption and accelerate training. The detailed experimental settings were as follows:

- Number of training epochs: 4 (num_train_epochs)
- LoRA application: Applied to all modules (lora_target_modules)
- Gradient checkpointing: Enabled to reduce memory usage (gradient_checkpointing)
- Batch size: 5
- Weight decay: 5 (weight_decay) to prevent overfitting
- Learning rate: 1e-4 (learning_rate) for stable fine-tuning
- Gradient accumulation steps: 16 (gradient_accumulation_steps) to simulate a larger batch size and improve training stability

This strategy ensures efficient fine-tuning and reliable task-specific performance while maintaining computational efficiency.

Table 2. The parameters of fine-tuning

Parameters	Value
CPU	Intel Xeon Gold 5218 64 cores
GPU	NVIDIA V100 Tensor Core 32Gb*2
Memory	256Gb
sft_type	'lora'
dtype	'fp16'
num_train_epochs	4
lora_target_modules	'ALL'
gradient_checkpointing	True
batch_size	5
weight_decay	0.1
learning_rate	1e-4
gradient_accumulation_steps	16
Others	Default

4. Experience

4.1. Assessment Methods

Aper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use. This study evaluates four open-source large language models—Qwen2, GLM4, Baichuan2, and LLaMA3.1—using identical parameter settings and fine-tuning methods. The information text generation task was assessed through three primary metrics:

Text Fluency (BLEU): The fluency of generated text was measured using BLEU, which evaluates the grammatical correctness and readability of the text by assessing its similarity to reference texts. Although BLEU is straightforward to implement, it overly depends on word order and exact matches, making it less effective at capturing semantic similarity. To address this limitation, the study incorporates perplexity (PPL), calculated using a pre-trained Chinese BERT model, to better evaluate the semantic coherence and fluency of the text.

$$\text{BLEU} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} PPL(S_i)}{100} \tag{1}$$

Text Novelty (NOV): Text novelty measures the degree of divergence between generated texts and the training data to ensure that the model does not merely replicate training content. This was computed by determining the Jaccard similarity between generated sentences and the training corpus, with the novelty score calculated as NOV=1−max (Jaccard similarity). An appropriate NOV score indicates the model's capability for creative text generation.

$$\text{NOV}(S_g) = 1 - \text{Max}\{\text{Jaccard}(S_g, C_i)\}_{,i=1\sim|c|} \tag{2}$$

Text Diversity (DIV): Text diversity assesses the variability in vocabulary and sentence structure, ensuring the model does not fall into repetitive patterns. It was evaluated by calculating the Jaccard similarity among sentences within the set of generated texts. Higher DIV scores reflect the model's ability to produce a wide range of expressions and content.

$$\text{DIV}(S_g) = 1 - \text{Max}\{\text{Jaccard}(S_g, S_i)\}_{,i=1\sim|s|,i\neq g} \tag{3}$$

These metrics provide a comprehensive framework to assess the quality, creativity, and variability of generated text, ensuring a robust evaluation of the models' performance in information text generation tasks. Equation 4 indicates the relevance of the generated false information to keywords, Equation 5 indicates the relevance of the generated rumor-refuting information to and false information, and Equations 6 and 7 measure the stability of the overall generation results of the model by statistically repeating the words in the generated sentences.

$$\text{Relevance}_{\text{fake}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{KeywordCount}(S_i)}{\text{TotalKeywords}} \tag{4}$$

$$\text{Relevance}_{\text{true}} = \frac{1}{N}\sum_{i=1}^{N} \text{Similarity}(S_i, S_{\text{FalseInfo}}) \tag{5}$$

$$\text{Repeat}(S_i) = \frac{1}{L}\sum_{j=1}^{L} \mathbb{I}\big(\text{Count}(w_j, S_i) > T\big) \tag{6}$$

$$\text{AnomalyRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\text{Repeat}(S_i) > 0) \tag{7}$$

4.2. Results

After fine-tuning and task-specific evaluation of the four models, the results are summarized in the following **Table 3**.

Qwen2 performs the best in terms of text fluency and anomaly rates for both tasks. It also has the shortest inference time. However, this shorter inference time is due to its generation of shorter text with fewer event details, and its GPU inference usage is relatively high.

GLM4 excels in generating longer texts with a higher degree of novelty in the content. However, it suffers from a higher anomaly rate and excessive repetition in responses, making the results difficult to control. Additionally, the inference time is considerably long.

Baichuan2 has lower parameter size and thus lower GPU utilization, resulting in shorter inference times. However, this comes at the cost of model performance, with lower fluency, novelty, diversity, and relevance in text generation compared to the other models.

LLaMA3.1 stands out in terms of diversity and relevance, generating richer content with a more balanced response. The generated text is appropriately long and includes relevant topics or titles, making it more suited to social media content. Although its inference performance is not the fastest, it offers a well-rounded performance. This model, fine-tuned with a Chinese network language instruction set using Direct Preference Optimization (DPO), improves its Chinese language capabilities. When compared to other domestic models, LLaMA3.1 demonstrates stronger stability in Chinese text generation. Additionally, with DPO applied to instruction sets from platforms like Zhihu and Qiangzhi Ba (an internet forum), LLaMA3.1 shows superior logical consistency in its generated content, producing more coherent and rational false information compared to the other models.

Table 3. Results of experimental assessment of different models.

Index	False information generation				Clarification information generation				Inference performance		
	BLEU	NOV	DIV	Relevance	BLEU	NOV	DIV	Relevance	Abnomal	GPU used	Time
Qwen2	0.9719	0.9625	0.9296	0.8562	0.9807	0.9546	0.9527	0.8032	0.0085	58%	4s
GLM4	0.9693	0.9714	0.9327	0.7938	0.9631	0.9520	0.9433	0.7548	0.0120	50%	13s
Baichuan2	0.9642	0.9556	0.8734	0.8235	0.9642	0.9475	0.8964	0.7207	0.0105	40%	5s
LLaMA3.1	0.9613	0.9601	0.9359	0.8924	0.9601	0.9588	0.9734	0.8236	0.0110	45%	6s

The fine-tuned model generation results are shown in the table below, where example prompts for natural disaster information and public health medical health information were used to query the four models. The LLaMA3.1 model generated responses of moderate length. In generating false information, it paid attention to the need for an inflammatory tone, making its responses more engaging and persuasive compared to the other models. In generating clarifying information, it provided further explanations of the reasons behind the information, guided people not to trust non-authoritative sources, and offered advice. Thus, the fine-tuned LLaMA3.1 model was chosen as the foundational model for subsequent agent construction in this study.

Table 4. Examples of information generation results for different models.

Parameters	False information generation	Clarification information generation
Prompt words	· Based on the keywords [Gong County, Yibin, Sichuan, a mountain landslide accident occurred, resulting in untrue deaths and missing persons], please generate a piece of social life false information for social media dissemination, use an inciting tone, and reflect negative emotions.	· Based on the keywords [Gong County, Yibin, Sichuan, a mountain landslide accident occurred, resulting in untrue deaths and missing persons], please generate a piece of social life false information for social media dissemination, use an inciting tone, and reflect negative emotions.
Qwen2	· URGENT NOTICE: A serious accident has occurred in Gongxian County, where a mountain collapse has caused a large number of casualties! According to the latest news, dozens of people have been tragically killed and many more are still on the brink of life and death. All residents are urged to remain vigilant and avoid traveling to the affected areas.	· Recently, there have been rumors circulating on the Internet that the use of specific biocide humidifiers can effectively reduce the risk of respiratory infections. However, we would like to remind you that there are no studies that have proven that the addition of specific biocides to humidifiers is effective in preventing or reducing respiratory diseases. Conversely, improper use can cause potential health hazards. So, don't believe such unsubstantiated claims and make sure to follow professional advice to maintain indoor air hygiene.
GLM4	· The shocking news has just come from the county: a sudden landslide has killed at least five people and left many missing. The local	· Please note that claims that chemicals can be added to improve indoor air quality or prevent the spread of disease are unscientific. In fact,

	government is working hard to search for survivors, but the situation is not encouraging. Please pay close attention to official information and take safety precautions.	adding any form of disinfectant or sanitizer to a humidifier not only does not effectively purify the air, but can instead produce harmful gases and increase health risks. The right thing to do is to clean the humidifier regularly and keep it clean to avoid the growth of bacteria and mold.
Baichuan2	· According to reports, in Yibin, Sichuan, the area of Jue County has recently encountered a catastrophic natural disaster - a sudden landslide! This unforeseen and extremely dangerous landslide has caused many residents to be in distress and the situation is extremely serious, please pay attention to the safety of your family members immediately and take self-protection measures to deal with this sudden disaster. We call on all sectors of society to pay attention to and pass on this important message reminder!	· According to reports, in Yibin, Sichuan, the area of Jue County has recently encountered a catastrophic natural disaster - a sudden landslide! This unforeseen and extremely dangerous landslide has caused many residents to be in distress and the situation is extremely serious, please pay attention to the safety of your family members immediately and take self-protection measures to deal with this sudden disaster. We call on all sectors of society to pay attention to and pass on this important message reminder!
LLaMA3.1	· Urgent warning! In Yibin Gong County, Sichuan Province, a shocking tragedy has occurred. As a result of severe landslides caused by a rare rainstorm that has hit the area for days on end, local residents are experiencing unprecedented panic and chaos, according to the latest news. We urge everyone to take immediate action to keep themselves and their families safe, and to retweet to alert all community members! #Emergency Notice #Landslide Warning	· Recently, the claim that adding so-called 'specific disinfectants' to humidifiers can improve the cleanliness of household air and reduce viral infections is completely misleading to public health. The truth is that the microorganisms in the water source are less harmful and do not need to add any additional chemicals to prevent the growth of viruses or bacteria; Conversely, the direct addition of such disinfectants without rigorous scientific validation can pose serious health risks. Please be wary of such sources of information that have not been verified by the official authorities and follow professional guidance for daily humidifier cleaning operations!

5. Summary

This paper centers around the generation of rumor and clarification information during emergencies leveraging large language models. In the context of the escalating spread of false information in online emergency public opinion and the impact of large language models, this research is highly significant.

In the methodology, it presents several innovative aspects. Firstly, a specific-scenario dataset is constructed to fine-tune open-source large language models. By gathering and processing data from authoritative platforms, a high-quality dataset with 1,715 pairs of false and clarification information is formed. Secondly, it better caters to emergency scenarios, real-world situations, and public opinion dissemination needs. This enables the generation of more practical false information and clarification texts that are more authentic, targeted, and effective, thus better adapting to the actual public opinion environment. Through the LoRA fine-tuning approach with a task-separated strategy, the models are optimized.

Regarding the evaluation, metrics such as BLEU, NOV, and DIV are utilized. After fine-tuning and assessment, Qwen2 demonstrates good performance in text fluency and anomaly rates yet generates shorter texts. GLM4 produces longer and novel texts but has higher anomaly rates. Baichuan2 has lower GPU utilization but relatively weaker performance. LLaMA3.1 excels in diversity and relevance and is selected as the foundation for subsequent agent construction.

However, the study also has its limitations. Although large language models can generate false and clarification information with strong logic and diversity, the consistency and stability of long text generation can be further enhanced. Future work could involve increasing domain-specific data training, improving data quality, and adopting new alignment methods like more refined RLHF (Reinforcement Learning from Human Feedback) to improve the logic and stability of model generation results. Overall, this research makes substantial contributions to understanding and resolving the issue of false information generation and clarification during emergencies, offering valuable methods and models for related research and practical applications.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China. Project number is 72274096.

References

[1] Sequence to Sequence Learning with Neural Networks. [EB/OL], 2014

[2] Qiu X, Sun T, Xu Y, et al. pre-trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.

[3] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[4] Kenton J D M-W C, Toutanova L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of naacL-HLT, 2019: 2.

[5] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[6] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67.

[7] Crothers E N, Japkowicz N, Viktor H L. Machine-generated text: A comprehensive survey of threat models and detection methods[J]. IEEE Access, 2023, 11: 70977-71002.

[8] Generative language models and automated influence operations: Emerging threats and potential mitigations. [EB/OL]. arXiv preprint arXiv:2301.04246, 2023

[9] FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models. [EB/OL]. arXiv preprint arxiv:2310.05046, 2024

[10] Yue Huang L. FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models[J].

[11] Sun Y, He J, Cui L, et al. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges[J]. arXiv preprint arXiv:2403.18249, 2024.

[12] Lai J, Yang X, Luo W, et al. RumorLLM: A Rumor Large Language Model-Based Fake-News-Detection Data-Augmentation Approach[J]. Applied Sciences, 2024, 14(8): 3532.

[13] Qwen2 technical report. [EB/OL]. arXiv preprint arXiv:2407.10671, 2024

[14] Baichuan 2: Open large-scale language models. [EB/OL]. arXiv preprint arXiv:2309.10305, 2023

[15] Chatglm: A family of large language models from glm-130b to glm-4 all tools. [EB/OL]. arXiv preprint arXiv:2406.12793, 2024

[16] The llama 3 herd of models. [EB/OL]. arXiv preprint arXiv:2407.21783, 2024

[17] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[J], 2016.

[18] Wang Y, Yang W, Ma F, et al. Weak supervision for fake news detection via reinforcement learning[C]. Proceedings of the AAAI conference on artificial intelligence, 2020: 516-523.

[19] Zhang X, Cao J, Li X, et al. Mining dual emotion for fake news detection[C]. Proceedings of the web conference 2021, 2021: 3465-3476.

[20] Nan Q, Cao J, Zhu Y, et al. MDFEND: Multi-domain fake news detection[C]. Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021: 3343-3347.